Overview of the ImageCLEFphoto 2007 photographic retrieval task

Michael Grubinger[1], Paul Clough[2], Allan Hanbury[3], Henning Müller[4]

[1] Victoria University, Melbourne, Australia
[2] Sheffield University, Sheffield, UK
[3] Vienna University of Technology, Vienna, Austria
[4] University and Hospitals of Geneva, Switzerland

Abstract

ImageCLEFphoto 2007 is the general photographic ad-hoc retrieval task of the *ImageCLEF 2007* evaluation campaign and provides both the resources and the framework necessary to perform comparative laboratory-style evaluation of visual information retrieval from generic photographic collections. In 2007, the evaluation objective concentrated on retrieval of lightly annotated images, a new challenge that attracted a large number of submissions: a total of 20 participating groups submitting a record number of 616 system runs. This paper summarises the components used in the benchmark, including the document collection, the search tasks, an analysis of the submissions from participating groups, and results.

The participants were provided with a subset of the *IAPR TC-12 Benchmark*: 20,000 colour photographs and four sets of semi-structured annotations in (1) English, (2) German, (3) Spanish and (4) one set whereby the annotation language had randomly been selected for each of the images. Unlike in 2006, the participants were not allowed to use the semantic description field in their retrieval approaches. The topics and relevance assessments from 2006 were reused (and updated) to facilitate the comparison of retrieval from fully and lightly annotated images.

Some of the findings for multilingual visual information retrieval from generic collections of lightly annotated photographs include: bilingual retrieval performs as well as monolingual retrieval; the choice of the query language is almost negligible as many of the short captions contain proper nouns; combining concept and content-based retrieval methods as well as using relevance feedback and/or query expansion techniques can significantly improve retrieval performance; and the retrieval results are similar to those in 2006, despite the limited image annotations in 2007.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Performance Evaluation, IAPR TC-12 Benchmark, Image Retrieval

1 Introduction

ImageCLEFphoto 2007 provides a system-centered evaluation for multilingual visual information retrieval from generic photographic collections (*i.e.* containing everyday real-world photographs akin to those that can frequently be found in private photographic collections).

1.1 Evaluation Scenario

The evaluation scenario is similar to the classic TREC[1] ad-hoc retrieval task: simulation of the situation in which a system knows the set of documents to be searched, but cannot anticipate the particular topic that will be investigated (*i.e.* topics are not known to the system in advance) [21]. The goal of the simulation is: given an alphanumeric statement (and/or sample images) describing a user information need, find as many relevant images as possible from the given collection (with the query language either being identical or different from that used to describe the images).

1.2 Evaluation Objective 2007

The objective of *ImageCLEFphoto 2007* comprised the evaluation of multilingual visual information retrieval from a generic collection of lightly annotated photographs (*i.e.* containing only short captions such as the title, location, date or additional notes, but without a semantic description of that particular photograph). This new challenge allows for the investigation of the following research questions:

- Are traditional text retrieval methods still applicable for such short captions?
- How significant is the choice of the retrieval language?
- How does the retrieval performance compare to retrieval from collections containing fully annotated images (*ImageCLEFphoto 2006*)?
- Has the general retrieval performance improved in comparison with retrieval from lightly annotated images (*ImageCLEFphoto 2006*)?

One major goal of *ImageCLEFphoto 2007* was to attract more content-based retrieval approaches as most of the retrieval approaches in previous years had predominately been concept-based. The reduced alphanumeric semantic information provided with the image collection should support this goal as content-based retrieval techniques become more significant with more and more reduced image captions.

2 Evaluation Architecture

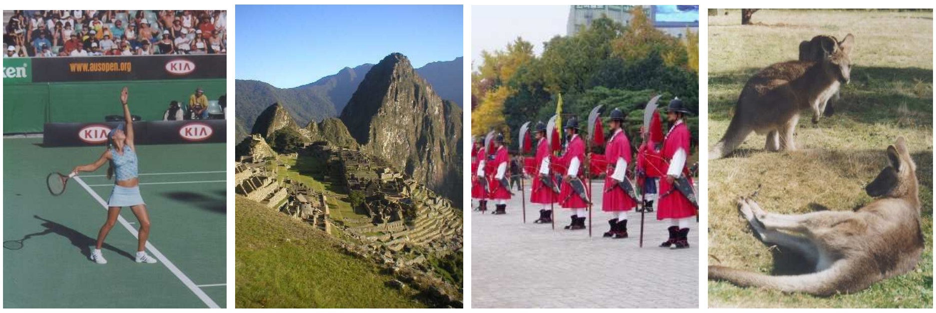
Similar to *ImageCLEFphoto 2006* [4], we generated a subset of the *IAPR TC-12 Benchmark* to provide the evaluation resources for *ImageCLEFphoto 2007*. This section provides more information on these individual components: the document collection, the query topics, relevance judgments and performance indicators. More information on the design and implementation of the *IAPR TC-12 Benchmark* itself, created under *Technical Committee 12* (TC-12) of the *International Association of Pattern Recognition* (IAPR[2]), can be found in [10].

2.1 Document Collection

The document collection of *IAPR TC-12 Benchmark* contains 20,000 colour photos taken from locations around the world and comprises a varying cross-section of still natural images. Figure 2.1 illustrates a number of sample images from a selection of categories.

[1] http://trec.nist.gov/

[2] http://www.iapr.org/



Sports. Landscapes. People. Animals.

Figure 1: Sample images from the *IAPR TC-12 collection*.

The majority of images have been provided by *viventura*[3], an independent travel company that organises adventure and language trips to South America. Travel guides accompany the tourists and maintain a daily online diary including photographs of trips made and general pictures of each location including accommodation, facilities and ongoing social projects. The remainder of the images have been collected by the first author over the past few years from personal experiences (*e.g.* holidays). The collection is publicly available for research purposes and, unlike many existing photographic collections used to evaluate image retrieval systems, this collection is very general in content with many different images of similar visual content, but varying illumination, viewing angle and background. This makes it a challenge for the successful application of techniques involving visual analysis.

Each image in the collection has a corresponding semi-structured caption consisting of the following seven fields: (1) a unique identifier, (2) a title, (3) a free-text description of the semantic and visual contents of the image, (4) notes for additional information, (5) the provider of the photo and fields describing (6) where and (7) when the photo was taken. Figure 2.1 shows a sample image with its corresponding English annotation.

```
<DOC>
<DOCNO>annotations/16/16019.eng</DOCNO>
<TITLE>Flamingo Beach</TITLE>
<DESCRIPTION>a photo of a brown sandy beach; the dark
  blue sea with small breaking waves behind it; a dark
  green palm tree in the foreground on the left; a blue
  sky with clouds on the horizon in the background;
</DESCRIPTION>
<NOTES>Original name in Portuguese: "Praia do Flamengo";
  Flamingo Beach is considered as one of the most
  beautiful beaches of Brazil;</NOTES>
<LOCATION>Salvador, Brazil</LOCATION>
<DATE>2 October 2004</DATE>
<IMAGE>images/16/16019.jpg</IMAGE>
<THUMBNAIL>thumbnails/16/16019.jpg</THUMBNAIL>
</DOC>
```

Figure 2: Sample image caption.

These annotations are stored in a database, allowing the creation of collection subsets with respect to a variety of particular parameters (*e.g.* which caption fields to use). Based on the feedback from participants of previous evaluation tasks, the following was provided for *ImageCLEFphoto 2007*:

- **Annotation language:** four sets of annotations in (1) English, (2) German, (3) Spanish and (4) one set whereby the annotation language was randomly selected for each of the images.

[3] `http://www.viventura.de/`

- **Caption fields:** only the fields for the *title*, *location*, *date* and additional *notes* were provided. Unlike 2006, the *description* field was not made available for retrieval to provide a more realistic evaluation scenario and to attract more visually oriented retrieval approaches.
- **Annotation completeness:** each image caption exhibited the same level of annotation completeness - there were no images without annotations as in 2006.

The participants were granted access to the data set on 16 April 2007 and had about three weeks to familiarise themselves with the new subset so that they could, for instance, adapt their existing retrieval scripts to the reduced multilingual annotations or to extract the visual and textual features of the images and their annotations in order to index the entire collection.

2.2 Query Topics

On 6 May 2007, the participants were given 60 query topics (see Table 1) representing typical search requests for the generic photographic collection of the *IAPR TC-12 Benchmark*.

ID	Topic Title	ID	Topic Title
1	accommodation with swimming pool	31	volcanos around Quito
2	church with more than two towers	32	photos of female guides
3	religious statue in the foreground	33	people on surfboards
4	group standing in front of mountain landscape in Patagonia	34	group pictures on a beach
5	animal swimming	35	bird flying
6	straight road in the USA	36	photos with Machu Picchu in the background
7	group standing in salt pan	37	sights along the Inca-Trail
8	host families posing for a photo	38	Machu Picchu and Huayna Picchu in bad weather
9	tourist accommodation near Lake Titicaca	39	people in bad weather
10	destinations in Venezuela	40	tourist destinations in bad weather
11	black and white photos of Russia	41	winter landscape in South America
12	people observing football match	42	pictures taken on Ayers Rock
13	exterior view of school building	43	sunset over water
14	scenes of footballers in action	44	mountains on mainland Australia
15	night shots of cathedrals	45	South American meat dishes
16	people in San Francisco	46	Asian women and/or girls
17	lighthouses at the sea	47	photos of heavy traffic in Asia
18	sport stadium outside Australia	48	vehicle in South Korea
19	exterior view of sport stadia	49	images of typical Australian animals
20	close-up photograph of an animal	50	indoor photos of churches or cathedrals
21	accommodation provided by host families	51	photos of goddaughters from Brazil
22	tennis player during rally	52	sports people with prizes
23	sport photos from California	53	views of walls with asymmetric stones
24	snowcapped buildings in Europe	54	famous television (and telecommunication) towers
25	people with a flag	55	drawings in Peruvian deserts
26	godson with baseball cap	56	photos of oxidised vehicles
27	motorcyclists racing at the Australian Motorcycle Grand Prix	57	photos of radio telescopes
28	cathedrals in Ecuador	58	seals near water
29	views of Sydney's world-famous landmarks	59	creative group pictures in Uyuni
30	room with more than two beds	60	salt heaps in salt pan

Table 1: *ImageCLEFphoto 2007* topics.

These topics had already been used in 2006, and we decided to reuse them to facilitate the objective comparison of retrieval from a generic collection of fully annotated (2006) and lightly annotated (2007) photographs. The creation of these topics had been based on several factors (see [9] for detailed information), including:

- the analysis of a log file from online-access to the image collection;
- knowledge of the contents of the image collection;
- various types of linguistic and pictorial attributes;
- the use of geographical constraints;
- the estimated difficulty of the topic.

Similar to TREC, the query topics were provided as structured statements of user needs which consist of a title (a short sentence or phrase describing the search request in a few words) and three sample images that are relevant to that search request. These images were removed from the test collection and did not form part of the ground-truth in 2007.

The topic titles were offered in 16 languages including English, German, Spanish, Italian, French, Portuguese, Chinese, Japanese, Russian, Polish, Swedish, Finnish, Norwegian, Danish, and Dutch, whereby all translations had been provided by at least one native speaker and verified by at least another native speaker. The participants only received the topic titles, but not the narrative descriptions to avoid misunderstandings as they had been misinterpreted by participants in the past (they only serve to unambiguously define what constitutes a relevant image or not).

2.3 Relevance Assessments

Relevance assessments were carried out by the two topic creators[4] using a custom-built online tool. The top 40 results from all submitted runs were used to create image pools giving an average of 2,299 images (max: 3237; min: 1513) to judge per topic.

The topic creators judged all images in the topic pools and also used interactive search and judge (ISJ) to supplement the pools with further relevant images. The assessments were based on a ternary classification scheme: (1) relevant, (2) partially relevant, and (3) not relevant. Based on these judgments, only those images judged relevant by both assessors were considered for the sets of relevant images (qrels).

Finally, these qrels were complemented with the relevant images found at *ImageCLEFphoto 2006* in order to avoid missing out on relevant images not found this year due to the reduced captions.

2.4 Result Generation

Once the relevance judgments were completed, we were able to evaluate the performance of the individual systems and approaches (the deadline for this result generation process was 15 July 2007). The results for submitted runs were computed using the latest version of trec_eval[5].

The submissions were evaluated using uninterpolated (arithmetic) *mean average precisions* (MAP) and *precision at rank 20* (P20) because most online image retrieval engines like *Google*, *Yahoo!* and *Altavista* display 20 images by default. Further measures considered include *geometric mean average precision* (GMAP) to test system robustness, and the *binary preference* (bpref) measure which is a good indicator for the completeness of relevance judgments.

3 Participation and Submission Overview

ImageCLEFphoto 2007 saw the registration of 32 groups (4 less than in 2006), with 20 of them eventually submitting a record number of 616 runs (all of which were evaluated). This is a drastic increase in comparison to previous years (12 groups submitting 157 runs in 2006, and 11 groups 349 runs in 2005 respectively).

Table 2 provides an overview of the participating groups, the corresponding number of submitted runs and the references of the working papers in which the participants describe their retrieval approaches. The 20 groups are from 20 different institutions in 16 countries, with one institution (Concordia University) sending two separate groups (CINDI, CLAC), while DCU and UTA joined forces and submitted as one participating group. New participants submitting in 2007 include Budapest, CLAC, UTA, NTU (Hongkong), ImpColl, INAOE, RUG, SIG and XRCE.

The increasing participation at *ImageCLEFphoto* might be an indicator for the growing need for evaluation of visual information retrieval from generic photographic collections and the global interest of researchers world-wide to participate in evaluation events such as *ImageCLEFphoto*.

[4] One of the topic generators is a member of the viventura travel company.

[5] `http://trec.nist.gov/trec_eval/trec_eval.7.3.tar.gz`

Group ID	Institution	Runs	Reference
Alicante	University of Alicante, Spain	6	[16]
Berkeley	University of California, Berkeley, USA	19	[14]
Budapest	Hungarian Academy of Sciences, Budapest, Hungary	11	[1]
CINDI	Concordia University, Montreal, Canada	5	[17]
CLAC	Concordia University, Montreal, Canada	6	[7]
CUT	Technical University Chemnitz, Germany	11	[22]
DCU-UTA	Dublin City University, Dublin, Ireland & University of Tampere, Finland	138	[13]
GE	University and Hospitals of Geneva, Switzerland	2	[23]
HongKong	Nanyang Technological University, Hong Kong	62	[11]
ImpColl	Imperial College, London, UK	5	[15]
INAOE	INAOE, Puebla, Mexico	115	[12]
IPAL	IPAL, Singapore	27	[8]
Miracle	Daedalus University, Madrid, Spain	153	[20]
NII	National Institute of Informatics, Tokyo, Japan	3	
RUG	University of Groningen, The Netherlands	4	[19]
RWTH	RWTH Aachen University, Germany	10	[5]
SIG	Universite Paul Sabatier, Toulouse, France	9	[18]
SINAI	University of Jaén, Jaén, Spain	15	[6]
Taiwan	National Taiwan University, Taipei, Taiwan	27	[2]
XRCE	Cross-Content Analytics, Meylan, France	8	[3]

Table 2: Participating groups.

Further, the number of runs per participating group has dramatically risen as well, with participants submitting an average of 30.8 runs in 2007 (13.1 runs in 2006). However, this may rather be attributed to the fact that four sets of annotations were offered (compared to two in 2007) and that the participants were allowed to submit as many system runs as they desired.

3.1 Submission Overview by Retrieval Dimensions

Overall, 616 runs were submitted and categorised with respect to the following dimensions: query and annotation language, run type (automatic or manual), use of relevance feedback or automatic query expansion, and modality (text only, image only or combined).

Dimension	Type	Data 2007	Data 2006	Total
Query Mode	bilingual	234 (8)	78 (2)	312 (8)
	monolingual	187 (17)	64 (2)	251 (18)
	visual	53 (12)		53 (12)
Annotation Language	English	271 (17)	137 (2)	408 (18)
	German	83 (7)	5 (1)	88 (8)
	Spanish	33 (7)		33 (7)
	Random	32 (2)		32 (2)
	none	52 (12)		52 (12)
Modality	Text Only	167 (15)	121 (2)	288 (15)
	Mixed (Text & Image)	255 (13)	21 (1)	276 (13)
	Image Only	52 (12)		52 (12)
Query Manipulation	none	148 (19)	131 (1)	279 (19)
	Relevance Feedback	204 (9)		204 (9)
	Query Expansion	76 (4)		76 (4)
	Feedback and Expansion	46 (5)	11 (1)	57 (6)
Run Type	Manual	19 (3)		19 (3)
	Automatic	455 (19)	142 (2)	597 (19)

Table 3: Submission overview by dimensions and data set used.

Table 3 provides an overview of all submitted runs according to these dimensions (with the number of groups in parenthesis). Most submissions (91.6%) used the image annotations, with 8 groups submitting a total of 312 bilingual runs and 18 groups a total of 251 monolingual runs; 15 groups experimented with purely concept-based (textual) approaches (288 runs), 13 groups investigated the combination of content-based (visual) and concept-based features (276 runs), while a total of 12 groups submitted 52 purely content-based runs, a dramatic increase in comparison with previous events (in 2006, only 3 groups had submitted a total of 12 visual runs). Furthermore, 53.4% of all retrieval approaches involved the use of image retrieval (31% in 2006).

Based on all submitted runs, 50.6% were bilingual (59% in 2006), 54.7% of runs used query expansion and pseudo-relevance feedback techniques (or both) to further improve retrieval results (46% in 2006), and most runs were automatic (*i.e.* involving no human intervention); only 3.1% of the runs submitted were manual.

Two participating groups made use of additional data (*i.e.* the description field and the qrels) from *ImageCLEFphoto 2006*. Although all these runs were evaluated (indicated by "Data 2006"), they were not considered for the system performance analysis and retrieval evaluation described in Section 4.

3.2 Submission Overview by Languages

Table 4 displays the number of runs (and participating groups in parenthesis) with respect to query and annotation languages. The majority of runs (66.2%) was concerned with retrieval from English annotations, with exactly half of them (33.1%) being monolingual experiments and all groups (except for GE and RUG) submitting at least one monolingual English run. Participants also showed increased interest in retrieval from German annotations; a total of eight groups submitted 88 runs (14.5% of total runs), 20.5% of them monolingual (compared with four groups submitting 18 runs in 2006). Seven groups made use of the new Spanish annotations (5.4% of total runs, 48.5% of them monolingual), while only two participants experimented with the annotations with a randomly selected language for each image (5.3%).

Query / Annotation	English	German	Spanish	Random	None	Total
English	204(18)	18 (5)	6 (3)	11 (2)		239(18)
German	31 (6)	31 (7)	1 (1)	11 (2)		74 (9)
Visual	1 (1)				52 (12)	53(12)
French	32 (7)	1 (1)	10 (2)			43 (7)
Spanish	20 (5)		16 (7)	2 (1)		38 (9)
Swedish	20 (3)	12 (1)				32 (3)
Simplified Chinese	24 (4)	1 (1)				25 (4)
Portuguese	19 (5)			2 (1)		21 (5)
Russian	17 (4)	1 (1)		2 (1)		20 (4)
Norwegian	6 (1)	12 (1)				18 (1)
Japanese	16 (3)					16 (3)
Italian	10 (4)			2 (1)		12 (4)
Danish		12 (1)				12 (1)
Dutch	4 (1)			2 (1)		6 (1)
Traditional Chinese	4 (1)					4 (1)
Total	408 (18)	88 (8)	33 (7)	32 (2)	52 (12)	616(20)

Table 4: Submission overview by query and annotation languages.

The expanded multilingual character of the evaluation environment also yielded an increased number of bilingual retrieval experiments: while only four query languages (French, Italian, Japanese, Chinese) had been used in 10 or more bilingual runs in 2006, a total of 13 languages were used to start retrieval approaches in 10 or more runs in 2007. The most popular languages this year were German (43 runs), French (43 runs) and English (35 runs). Surprisingly, 26.5% of the bilingual experiments used a Scandinavian language to start the retrieval approach: Swedish (32 runs), Norwegian (18 runs) and Danish (12 runs) – none of these languages had been used in 2006. It is also interesting to note that Asian languages (18.6% of bilingual runs) were almost exclusively used for retrieval from English annotations (only one run experimented with the German annotations), which might indicate a lack of translation resources from Asian to European languages other than English.

4 Results

This section provides an overview of the system results with respect to query and annotation languages as well as other submission dimensions such as query mode, retrieval modality and the involvement of relevance feedback or query expansion techniques.

Although the description fields were not provided with the image annotations, the absolute retrieval results achieved by the systems were not much lower compared to those in 2006 when the entire annotation was used. We attribute this to the fact that more than 50% of the groups had participated at ImageCLEF before, improved retrieval algorithms (not only of returning participants), and the increased use of content-based retrieval approaches.

4.1 Results by Language

Table 5 shows the runs which achieved the highest MAP for each language pair (ranked by descending order of MAP scores). Of these runs, 90.6% use query expansion or (pseudo) relevance feedback, and 78.1% use both visual and textual features for retrieval. It is noticeable that submissions from CUT, DCU, NTU (Taiwan) and INAOE dominate the results (see participants' workshop papers for further information about their runs). As in previous years, the highest English monolingual run slightly outperforms the highest German and Spanish monolingual runs (MAPs are 22.9% and 12.1% lower).

Language (Annotation)	Group	Run ID	MAP	P20	GMAP	bpref
English (English)	CUT	cut-EN2EN-F50	0.3175	0.4592	0.2984	0.1615
German (English)	XRCE	DE-EN-AUTO-FB-TXTIMG_MPRF	0.2899	0.3883	0.2684	0.1564
Portuguese (English)	Taiwan	NTU-PT-EN-AUTO-FBQE-TXTIMG	0.2820	0.3883	0.2655	0.1270
Spanish (English)	Taiwan	NTU-ES-EN-AUTO-FBQE-TXTIMG	0.2785	0.3833	0.2593	0.1281
Russian (English)	Taiwan	NTU-RU-EN-AUTO-FBQE-TXTIMG	0.2731	0.3825	0.2561	0.1146
Italian (English)	Taiwan	NTU-IT-EN-AUTO-FBQE-TXTIMG	0.2705	0.3842	0.2572	0.1138
S. Chinese (English)	CUT	cut-ZHS2EN-F20	0.2690	0.4042	0.2438	0.0982
French (English)	Taiwan	NTU-FR-EN-AUTO-FBQE-TXTIMG	0.2669	0.3742	0.2480	0.1151
T. Chinese (English)	Taiwan	NTU-ZHT-EN-AUTO-FBQE-TXTIMG	0.2565	0.3600	0.2404	0.0890
Japanese (English)	Taiwan	NTU-JA-EN-AUTO-FBQE-TXTIMG	0.2551	0.3675	0.2410	0.0937
Dutch (English)	INAOE	INAOE-NL-EN-NaiveWBQE-IMFB	0.1986	0.2917	0.1910	0.0376
Swedish (English)	INAOE	INAOE-SV-EN-NaiveWBQE-IMFB	0.1986	0.2917	0.1910	0.0376
Visual (English)	INAOE	INAOE-VISUAL-EN-AN_EXP_3	0.1925	0.2942	0.1921	0.0390
Norwegian (English)	DCU	NO-EN-Mix-sgramRF-dyn-equal-fire	0.1650	0.2750	0.1735	0.0573
German (German)	Taiwan	NTU-DE-DE-AUTO-FBQE-TXTIMG	0.2449	0.3792	0.2386	0.1080
English (German)	XRCE	EN-DE-AUTO-FB-TXTIMG_MPRF_FLR	0.2776	0.3617	0.2496	0.1121
Swedish (German)	DCU	SW-DE-Mix-dictRF-dyn-equal-fire	0.1788	0.2942	0.1802	0.0707
Danish (German)	DCU	DA-DE-Mix-dictRF-dyn-equal-fire	0.1730	0.2942	0.1759	0.0733
French (German)	CUT	cut-FR2DE-F20	0.1640	0.2367	0.1442	0.0039
Norwegian (German)	DCU	NO-DE-Mix-dictRF-dyn-equal-fire	0.1667	0.2700	0.1653	0.0701
Spanish (Spanish)	Taiwan	NTU-ES-ES-AUTO-FBQE-TXTIMG	0.2792	0.3975	0.2693	0.1128
English (Spanish)	CUT	cut-EN2ES-F20	0.2770	0.3767	0.2470	0.1054
German (Spanish)	Berkeley	Berk-DE-ES-AUTO-FB-TXT	0.0910	0.1217	0.0717	0.0080
English (Random)	DCU	EN-RND-Mix-sgramRF-dyn-equal-fire	0.1678	0.2850	0.1751	0.0683
German (Random)	DCU	DE-RND-Mix-sgram-dyn-equal-fire	0.1572	0.2817	0.1669	0.0644
French (Random)	DCU	FR-RND-Mix-sgram-dyn-equal-fire	0.1409	0.2642	0.1476	0.0593
Spanish (Random)	INAOE	INAOE-ES-RND-NaiveQE-IMFB	0.1243	0.2275	0.1355	0.0266
Dutch (Random)	INAOE	INAOE-NL-RND-NaiveQE	0.0828	0.1558	0.0941	0.0114
Italian (Random)	INAOE	INAOE-IT-RND-NaiveQE	0.0798	0.1442	0.0864	0.0181
Russian (Random)	INAOE	INAOE-RU-RND-NaiveQE	0.0763	0.1358	0.0848	0.0174
Portuguese (Random)	INAOE	INAOE-PT-RND-NaiveQE	0.0296	0.0425	0.0317	0.0006
Visual	XRCE	AUTO-NOFB-IMG_COMBFK	0.1890	0.3517	0.2009	0.1016

Table 5: Systems with highest MAP for each language.

The highest bilingual to English run (German – English) performed with a MAP of 91.3% of the highest monolingual run MAP, with the highest bilingual run in most other query languages such as Portuguese, Spanish, Russian, Italian, Chinese, French and Japanese all exhibiting at least 80% of that highest monolingual English run. Hence, there is no longer much difference between monolingual and bilingual retrieval, indicating a significant progress of the translation and retrieval methods using these languages. Moreover, the highest bilingual to Spanish run (English – Spanish) had a MAP of 99.2% of the highest monolingual Spanish run, while the highest bilingual to German run (English – German) even outperformed the highest German monolingual run MAP by 13.3%.

4.2 Results by Query Mode

This trend is not only true for the highest runs per language pair, but also for all submissions and across several performance indicators. Table 6 illustrates the average scores across all system runs (and the standard deviations in parenthesis) with respect to monolingual, bilingual and purely visual retrieval.

Query Mode	MAP	P(20)	BPREF	GMAP
Monolingual	0.1384 (0.0696)	0.1920 (0.1023)	0.1320 (0.0664)	0.0375 (0.0362)
Bilingual	0.1364 (0.0563)	0.1994 (0.0880)	0.1357 (0.0542)	0.0370 (0.0268)
Visual	0.0681 (0.0385)	0.1568 (0.0691)	0.0800 (0.0387)	0.0219 (0.0187)

Table 6: Results by query mode.

Again, monolingual and bilingual retrieval are almost identical, and so are the average results for monolingual Spanish, English and German retrieval (see Table 7): Spanish shows the highest average MAP and BPREF values, while German exhibits the highest average for P(20) and English for GMAP.

Annotation	MAP	P(20)	BPREF	GMAP
Spanish	0.1450 (0.0593)	0.1947 (0.0896)	0.1342 (0.0563)	0.0362 (0.0343)
English	0.1388 (0.0753)	0.1900 (0.1076)	0.1317 (0.0709)	0.0384 (0.0375)
German	0.1331 (0.0427)	0.2001 (0.0828)	0.1321 (0.0475)	0.0339 (0.0313)

Table 7: Monolingual results by annotation language.

Across all submissions, the average values for bilingual retrieval from English and German annotations are even slightly higher than those for monolingual retrieval (see Table 8), while bilingual retrieval from Spanish annotations and from annotations with a randomly selected language does not lag far behind.

Annotation	MAP	P(20)	BPREF	GMAP
English	0.1497 (0.0551)	0.2037 (0.0885)	0.1428 (0.0536)	0.0374 (0.0292)
German	0.1384 (0.0400)	0.2174 (0.0748)	0.1452 (0.0398)	0.0445 (0.0206)
Spanish	0.1171 (0.0787)	0.1758 (0.1081)	0.1083 (0.0703)	0.0273 (0.0368)
Random	0.0992 (0.0475)	0.1691 (0.0840)	0.1083 (0.0517)	0.0283 (0.0213)
None	0.0681 (0.0385)	0.1568 (0.0691)	0.0800 (0.0387)	0.0219 (0.0187)

Table 8: Bilingual results by annotation language.

These results indicate that the query language does not play a major factor for visual information retrieval for lightly annotated images. We attribute this (1) to the high quality of the state-of-the-art translation techniques, (2) to the fact that such translations implicitly expand the query terms (similar to query expansion using a thesaurus) and (3) to the short image captions used (as many of them are proper nouns which are often not even translated).

4.3 Results by Retrieval Modality

In 2006, the system results had shown that combining visual features from the image and semantic knowledge derived from the captions offered optimum performance for retrieval from a generic photographic collection with fully annotated images.

As indicated in Table 9, the results of *ImageCLEFphoto 2007* show that this also applies for retrieval from generic photographic collections with lightly annotated images: on average, combining visual features from the image and semantic information from the annotations gave a 24% improvement of the MAP over retrieval based solely on text.

Purely content-based approaches still lag behind, but the average MAP for retrieval solely based on image features shows an improvement of 65.8% compared to the average MAP in 2006.

Modality	MAP	P(20)	BPREF	GMAP
Mixed	0.1487 (0.0655)	0.2251 (0.0968)	0.2026 (0.0808)	0.0498 (0.0313)
Text Only	0.1199 (0.0404)	0.1519 (0.0509)	0.1408 (0.0447)	0.0180 (0.0180)
Image Only	0.0681 (0.0385)	0.1568 (0.0691)	0.0800 (0.0387)	0.0219 (0.0187)

Table 9: Results by retrieval modality.

4.4 Results by Feedback and/or Query Expansion

Table 10 illustrates the average scores across all systems runs (and the standard deviations in parenthesis) with respect to the use of query expansion or relevance feedback techniques.

Technique	MAP	P(20)	BPREF	GMAP
None	0.1094 (0.0518)	0.1779 (0.0748)	0.1100 (0.0473)	0.0272 (0.0236)
Query Expansion	0.1117 (0.0396)	0.1575 (0.0528)	0.1056 (0.0355)	0.0242 (0.0191)
Relevance Feedback	0.1312 (0.0547)	0.1849 (0.0844)	0.1315 (0.0535)	0.0374 (0.0265)
Expansion & Feedback	0.2182 (0.0620)	0.3236 (0.0760)	0.2090 (0.0525)	0.0726 (0.0465)

Table 10: Results by feedback or query expansion.

While the use of query expansion does not necessarily seem to dramatically improve retrieval results for retrieval from lightly annotated images (average MAP only 2.1% higher), relevance feedback (typically in the form of query expansion based on pseudo relevance feedback) appeared to work well on short captions (average MAP 19.9% higher), with a combination of query expansion and relevance feedback techniques yielding results almost twice as good as without any of these techniques (average MAP 99.5% higher).

4.5 Results by Run Type

Table 11 shows the average scores across all systems runs (and the standard deviations in parenthesis) with respect to the run type. Unsurprisingly, MAP results of manual approaches are, on average, 58.6% higher than purely automatic runs — this trend seems to be true for both fully annotated and lightly annotated images.

Technique	MAP	P(20)	BPREF	GMAP
Manual	0.2010 (0.0811)	0.3016 (0.1156)	0.1886 (0.0742)	0.0656 (0.0512)
Automatic	0.1267 (0.0579)	0.1872 (0.0838)	0.1256 (0.0545)	0.0343 (0.0285)

Table 11: Results by run type.

5 Conclusion

This paper reported on *ImageCLEFphoto 2007*, the general photographic ad-hoc retrieval task of the *ImageCLEF 2007* evaluation campaign. Its evaluation objective concentrated on visual information retrieval from generic collections of lightly annotated images, a new challenge that attracted a large number of submissions: 20 participating groups submitted a total of 616 system runs.

The participants were provided with a subset of the *IAPR TC-12 Benchmark*: 20,000 colour photographs and four sets of semi-structured annotations in (1) English, (2) German, (3) Spanish and (4) one set whereby the annotation language was randomly selected for each of the images. Unlike in 2006, the participants were not allowed to use the semantic description field in their retrieval approaches. The topics and relevance assessments from 2006 were reused (and updated) to facilitate the comparison of retrieval from fully and lightly annotated images.

The nature of the task also attracted a larger number of participants experimenting with content-based retrieval techniques, and hence the retrieval results were similar to those in 2006, despite the limited image annotations in 2007. Other findings for multilingual visual information retrieval from generic collections of lightly annotated photographs include:

- bilingual retrieval performs as well as monolingual retrieval;
- the choice of the query language is almost negligible as many of the short captions contain proper nouns;
- combining concept and content-based retrieval methods as well as using relevance feedback and/or query expansion techniques can significantly improve retrieval performance;

ImageCLEFphoto will continue to provide resources to the retrieval and computational vision communities to facilitate standardised laboratory-style testing of image retrieval systems. While these resources have predominately been used by systems applying a concept-based retrieval approach thus far, the rapid increase of participants using content-based retrieval techniques at *ImageCLEFphoto* calls for a more suitable evaluation environment for visual approaches (*e.g.* the preparation of training data). For *ImageCLEFphoto 2008*, we are planning to create new topics and will therefore be able to provide this year's topics and qrels as training data for next year.

References

[1] András Benczúr, István Biró, Mátyás Brendel, Károly Csalogány, Daróczy Bálint, and Dávid Siklósi. Cross-modal retrieval by text and image feature biclustering. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[2] Yih-Chen Chang and Hsin-Hsi Chen. Experiment for Using Web Information to do Query and Document Expansion. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[3] Stephane Clinchant, Jean-Michel Renders, and Gabriela Csurka. XRCE's Participation to ImageCLEFphoto 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[4] Paul David Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, and Henning Müller. Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In *Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).

[5] Thomas Deselaers, Tobias Gass, Tobias Weyand, and Hermann Ney. FIRE in ImageCLEF 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[6] M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T Martín-Valvidia, A. Montejo-Raez, and L.A. Ureña López. SINAI at ImageCLEF 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[7] Osama El Demerdash, Leila Kosseim, and Sabine Bergler. Experiments with Clustering the Collection at ImageCLEF 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[8] Sheng Gao, Jean-Pierre Chevallet, Thi Hoang Diem Le, Trong Ton Pham, and Joo Hwee Lim. IPAL at ImageCLEF 2007 Mixing Features, Models and Knowledge. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[9] Michael Grubinger. On the Creation of Query Topics for ImageCLEFphoto. In *Third MUSCLE / ImageCLEF Workshop on Image and Video Retrieval Evaluation*, Budapest, Hungary, September 2007.

[10] Michael Grubinger, Paul David Clough, Henning Müller, and Thomas Deselears. The IAPR–TC12 Benchmark: A New Evaluation Resource for Visual Information Systems. In *International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC'06*, pages 13–23, Genoa, Italy, May 22 2006.

[11] Steven C. H. Hoi. Cross-Language and Cross-Media Image Retrieval: An Empirical Study at ImageCLEF 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[12] Hugo Jair Escalante, Carlos A. Hernández, Aurelio López, Heidy M. Marín, Manuel Montes y Gómez, Eduardo Morales, Luis E. Sucar, and Luis Villaseñor. TIA-INAOE's Participation at ImageCLEF 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[13] A. Järvelin, P. Wilkins, T. Adamek, E. Airio, G. Jones, E. Sormunen, and A.F. Smeaton. DCU and UTA at Photographic ImageCLEF 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[14] Ray R. Larson. Linked Relevance Feedback for the ImageCLEF Photo Task. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[15] João Magalhães, Simon Overell, and Stefan Rüger. Exploring Image, Text and Geographic Evidences in ImageCLEF 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[16] Sergio Navarro, Fernando Llopis, Rafael Muñoz Guillena, and Elisa Noguera. Information retrieval of visual descriptions with IR-n system based on passages. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[17] M.M. Rahman, Bipin C. Desai, and Prabir Bhattacharya. Multi-Modal Interactive Approach to ImageCLEF 2007 Photographic and Medical Retrieval by CINDI. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[18] Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem. Using pseudo-relevance feedback to improve image retrieval results. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[19] Florence Tushabe and Michael H. F. Wilkinson. Content-Based Image Retrieval Using Shape-Size Pattern Spectra. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[20] Julio Villena-Román, Sara Lana-Serrano, José-Luis Martínez-Fernández, and José Carlos González-Cristóbal. MIRACLE at ImageCLEFphoto 2007: Evaluation of Merging Strategies for Multilingual and Multimedia Information Retrieval. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[21] Ellen M. Voorhees and Donna Harmann. Overview of the Seventh Text REtrieval Conference (TREC–7). In *The Seventh Text Retrieval Conference*, pages 1–23, Gaithersburg, MD, USA, November 1998.

[22] Thomas Wilhelm, Jens Kürsten, and Maximilian Eibl. Experiments for the ImageCLEF 2007 Photographic Retrieval Task. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[23] Xin Zhou, Julien Gobeill, Patrick Ruch, and Henning Müller. University and Hospitals of Geneva at ImageCLEF 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.